Rudolf Rosa, David Mareček, Aleš Tamchyna
{rosa,marecek,tamchyna}@ufal.mff.cuni.cz

Deepfix:

Statistical Post-editing of Statistical Machine Translation Using Deep Syntactic Analysis

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

ACL SRW, Sofia, 6th August 2013
Motivation

Source text in English:

EU criticizes not only the Greek government.
Motivation

- Source text in English:
  
  *EU criticizes not only the Greek government*

- Google Translate to Czech (6th Aug 2013):
  
  *EU kritizuje nejen řecká vláda*
Motivation

- Source text in English:
  
  *EU criticizes not only the Greek government*

- Google Translate to Czech (6\textsuperscript{th} Aug 2013):
  
  *EU kritizuje nejen řecká vláda*

  - *Not only the Greek government criticizes EU*
Motivation

- Source text in English:
  
  EU criticizes *not only* the Greek government

- Google Translate to Czech (6th Aug 2013):
  
  *EU kritizuje nejen řecká vláda* nominative (subject)
  
  - *Not only the Greek government* criticizes EU

- Post-editation by Deepfix:
  
  *EU kritizuje nejen řeckou vládu* accusative (object)
  
  - *EU criticizes not only the Greek government*
Outline

1. Problem definition
2. Sentence analysis
3. Sentence post-editing
4. Results
Outline

1. Problem definition
   - Errors in valency in SMT outputs

2. Sentence analysis

3. Sentence post-editing

4. Results
1. Problem definition
   - Errors in valency in SMT outputs

2. Sentence analysis (DEEP)
   - Deep dependency parsing

3. Sentence post-editing

4. Results
Outline

1. Problem definition
   ➔ Errors in valency in SMT outputs
2. Sentence analysis (DEEP)
   ➔ Deep dependency parsing
3. Sentence post-editing (FIX)
   ➔ Statistical model of valency
4. Results
Outline

1. Problem definition
   - Errors in valency in SMT outputs

2. Sentence analysis (DEEP)
   - Deep dependency parsing

3. Sentence post-editing (FIX)
   - Statistical model of valency

4. Results
   - Automatic & manual evaluation of Deepfix
Subject – object dichotomy

- **English:** position (left/right constituent)
  - Subject criticize Object

- **Czech:** morphological case (nominative/other); word order relatively free
  - Subject nominative *kritizovat* Object accusative
  - Object accusative *kritizovat* Subject nominative
  - Subject nominative Object accusative *kritizovat*
  - Object accusative Subject nominative *kritizovat*
Valency of criticize (kritizovat)

- example sentence
  - EU<sub>subject</sub> criticizing not only the Greek government<sub>object</sub>
  - EU<sub>nominative</sub> kritizuje nejen řeckou vládu<sub>accusative</sub>
Valency of *criticize* (*kritizovat*)

- example sentence
  - *EU* *criticizes not only the Greek government*
  - *EU* *kritizuje nejen řeckou vládu*

- a valency frame of a verb
  - subject *criticize* object
  - nominative *kritizovat* accusative
Valency of *criticize* (**kritizovat**)

- example sentence
  - EU \textbf{criticizes} not only the Greek government
  - EU \textbf{kritizuje} nejen řeckou vládu

- a valency frame of a verb
  - subject 
  - nominative
  - nominative
  - accusative

(position) (cases)
Valency of \textit{criticize} (\textit{kritizovat})

- example sentence
  - \textit{EU} \textsubscript{subject} \textit{criticizes not only} the \textit{Greek government} \textsubscript{object}
  - \textit{EU} \textsubscript{nominative} \textit{kritizuje nejen řeckou vládu} \textsubscript{accusative}

- a valency frame of a verb
  - subject \textit{criticize} object \textit{(position)}
  - nominative \textit{kritizovat} accusative \textit{(cases)}

- decomposition into head-argument pairs
  - \textit{(to criticize, government)} ~ \textit{(kritizovat, vládu)}
Sentence analysis (DEEP)

- tagging & lemmatization
  - combination of rule-based and statistical approach
- word-alignment
  - unsupervised methods (Giza++)
- dependency parsing
  - statistical, trained on manually created treebanks
  - parser adapted for parsing of SMT outputs
- induction of deep structure (tectogrammar)
  - rule-based
Deep syntactic dependency trees

EU criticizes the Greek government

EU kritizuje řecká vláda
Deep syntactic dependency trees

EU criticizes the Greek government

EU kritizuje řecká vláda

criticize
government

EU

Greek

řecká

vláda
EU criticizes the Greek government

EU kritizuje řecká vláda
(head, arg) pair identification

EU criticizes the Greek government

EU kritizuje řecká vláda
Valency models (FIX)

- \( P(\text{arg}_{\text{case}} \mid \text{head}_{\text{lemma}}, \text{English}_{\text{arg}}_{\text{case}}) \)
- \( P(\text{arg}_{\text{case}} \mid \text{head}_{\text{lemma}}, \text{English}_{\text{arg}}_{\text{case}}, \text{arg}_{\text{lemma}}) \)
- estimated from CzEng 1.0 (15M parallel stcs)
Argument case probabilities

- \( P(\text{nominative} \mid \text{kritizovat}, \text{object}) = 0.03 \)
- \( P(\text{accusative} \mid \text{kritizovat}, \text{object}) = 0.80 \)
Argument case probabilities

- \( P(\text{nominative} \mid kritizovat, \text{object}) = 0.03 \)
- \( P(\text{accusative} \mid kritizovat, \text{object}) = 0.80 \)
- threshold: 0.55
Argument case correction

- $P(\text{nominative} \mid kritizovat, \text{object}) = 0.03$
- $P(\text{accusative} \mid kritizovat, \text{object}) = 0.80$
- threshold: 0.55
Sentence correction

- Statitical machine translation output:

  \[ EU \text{ kritizuje nejen } \overset{\text{nominative}}{\text{řecká}} \text{ nominative } \overset{\text{nominative}}{\text{vláda}} \text{ nominative} \]

- Not only the Greek government criticizes EU
Sentence correction

- Statistical machine translation output:

\[ EU \text{kritizuje nejen } \text{řecká}^{\text{nominative}} \text{ vláda}^{\text{nominative}} \]

- Not only the Greek government criticizes EU

- Valency model correction:

\[ EU \text{kritizuje nejen } \text{řecká}^{\text{nominative}} \text{ vládu}^{\text{accusative}} \]
Sentence correction

- Statistical machine translation output:

  EU kritizuje nejen řecká nominative vláda nominative

  - Not only the Greek government criticizes EU

- Valency model correction:

  EU kritizuje nejen řecká nominative vládu accusative

- Agreement enforcement:

  EU kritizuje nejen řekou accusative vládu accusative

  - EU criticizes not only the Greek government
Automatic evaluation (BLEU)

Rosa et al. – Deepfix: Statistical Post-editing of Statistical Machine Translation Using Deep Syntactic Analysis
Automatic evaluation (NIST)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SMT output</th>
<th>after Deepfix</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT10 (devel)</td>
<td>5.44</td>
<td>5.47</td>
</tr>
<tr>
<td>WMT11</td>
<td>5.73</td>
<td>5.74</td>
</tr>
<tr>
<td>WMT12</td>
<td>5.26</td>
<td>5.28</td>
</tr>
</tbody>
</table>
Automatic evaluation (1-PER)

<table>
<thead>
<tr>
<th>Year</th>
<th>SMT output</th>
<th>After Deepfix</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT10 (devel)</td>
<td>41.56</td>
<td>41.74</td>
</tr>
<tr>
<td>WMT11</td>
<td>42.83</td>
<td>42.91</td>
</tr>
<tr>
<td>WMT12</td>
<td>39.96</td>
<td>40.09</td>
</tr>
</tbody>
</table>
Manual evaluation (changed stcs)

- **Improvement**: 321 (56%)
- **Degradation**: 135 (24%)
- **Indefinite**: 113 (20%)
Conclusion

- address valency errors
  - statistical post-editing of SMT
- identify head-argument pairs (DEEP)
  - deep syntactic analysis
- find the best case for the arguments (FIX)
  - statistical valency model
- obtain slight improvement of translation quality
  - indicated by automatic evaluation
  - confirmed by manual evaluation
Future work

- explore existing valency lexicons
- more intricate modelling
  - combine more models
  - machine learning (now thresholds semi-manual, and overfitted to development data)
- further adapt underlying NLP tools (tagger)
- extend to other language pairs
Thank you for your attention

Rudolf Rosa, David Mareček, Aleš Tamchyna
{rosa,marecek,tamchyna}@ufal.mff.cuni.cz

Deepfix:
Statistical Post-editing
of Statistical Machine Translation
Using Deep Syntactic Analysis

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

For this presentation and other information, please visit:

http://ufal.mff.cuni.cz/~rosa/